Chapter 14.1 – Multiple Regression

Introduction

- Chapter 14 extends the linear model data addressed in Chapter 13.
- Now allow for more than one explanatory (x) variable.
- Much of the material covered in Chapter 13 carries over to this topic, but the addition of multiple explanatory variables introduces a much greater level of complexity in terms of both calculation and analysis of the results.

Multiple regression analysis by the method of OLS is probably the most commonly used statistical procedure in the social sciences. The OLS multiple regression model can be usefully applied to many problems, and has been modified and extended in a large number of ways. It is a powerful, and relatively robust tool, but is often mis-used, or used with insufficient caution.

A multiple regression analysis really consists of three main phases:

First, the data must be examined in univariate and bivariate ways to make sure that the most basic assumptions underlying OLS linear regression are being observed.

Second, a model must be selected, fit, and the parameters and goodness of fit assessed.

Third, before we take the results seriously, additional diagnostics should be performed on the residuals. Actually, residuals analysis must be conducted and produce satisfactory results before a candidate model can really be accepted for interpretation. In practice, however, full diagnosis of residuals is usually put off until we are reasonably sure about the basic model specification.

I. Population multiple regression equation

The simple linear model describes the linear relationship between a response variable, *y*, and a single explanatory variable, *x*. Now we will use a model in which the response variable will depend on *k* explanatory variables. As before, the model assumes than the mean of the response variable, μ_y, is a linear function of the explanatory variables.

Population multiple regression equation: $\mu_y = +\alpha\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$

Remember that this equation is describing the mean of the responses – as before, the actual responses will vary about this mean.

- **Multiple linear regression model**

The *statistical model for multiple linear regression* is

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

for $i = 1, 2, \ldots n$

The deviations ε_i are assumed to be independent and normally distributed with mean = 0 and standard deviation = σ. In other words, they are thought to be a simple random sample from the $N(0, \sigma)$ distribution.

The parameters of the multiple regression model are $\alpha, \beta_1, \beta_{2,\ldots}\beta_p$, and σ .

An Example:

```
use http://www.gseis.ucla.edu/courses/data/gpa

correlate gpa greq grev mat ar, means
(obs=30)

 Variable |         Mean    Std. Dev.          Min          Max
----------+--------------------------------------------------
      gpa |     3.313333     .5998467          2.5          4.3
     greq |     565.3333     48.61767          500          655
     grev |     575.3333     83.03441          480          720
      mat |           67     9.247553           55           85
       ar |     3.566667     .8384441          2.5            5


         |      gpa     greq     grev      mat       ar
---------+-------------------------------------------
     gpa |   1.0000
    greq |   0.6111   1.0000
    grev |   0.5815   0.4681   1.0000
     mat |   0.6042   0.2669   0.4257   1.0000
      ar |   0.6207   0.5078   0.4053   0.5248   1.0000

. regress gpa greq grev mat ar

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  4,    25) =   11.13
       Model |  6.68313355     4  1.67078339           Prob > F      =  0.0000
    Residual |   3.7515331    25  .150061324           R-squared     =  0.6405
-------------+------------------------------           Adj R-squared =  0.5829
       Total |  10.4346666    29  .359816091           Root MSE      =  .38738

------------------------------------------------------------------------------
         gpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        greq |   .0039983   .0018307     2.18   0.039      .000228    .0077686
        grev |   .0015236   .0010502     1.45   0.159    -.0006392    .0036865
         mat |   .0208961   .0095488     2.19   0.038     .0012299    .0405623
          ar |   .1442335   .1130013     1.28   0.214    -.0884969     .376964
       _cons |  -1.738107   .9507399    -1.83   0.079    -3.696192    .2199789
------------------------------------------------------------------------------
```

II. Some Preliminary Conclusions

An increase in the GRE quantitative score by one point will increase GPA by .0039983. An increase in the Miller Analogies Test by one point will increase GPA by .0208961 for a fixed level of GRE quantitative. The other variables (GRE Verbal and Average Teacher Rating) in the model are not statistically significant and the suggestion is that it appears that they have no effect on GPA. Note the negative intercept, this suggests (though nonsensical) that if all the other variables were zero, GPA would be –1.738107. This is extrapolation (WHY?).

III. Residuals Again (OPTIONAL)

As in Chapter 13, we use the residuals to create an estimate for σ:

$$s^2 = \frac{\sum e_i^2}{n-p-1} = \frac{\sum (y_i - \hat{y}_i)^2}{n-p-1}$$

and s is just the square root of s^2, the **degrees of freedom** for s^2 is n-p-1 and s will be our estimator of σ

In Stata, to get predicted values of y, first run the regression, then immediately issue the command

```
predict p
```

```
(option xb assumed; fitted values)
```

to generate residuals issue the command:

```
predict r, rstu
```

to get a sense of how well you are doing, it's worth it to examine the residuals and the predicted values:

```
list gpa p r
```

	gpa	p	r
1.	3.2	3.331268	-.3830797
2.	4.1	3.813245	.7844458
3.	3	2.791182	.5684165
4.	2.6	2.829665	-.6148894
5.	3.7	3.174036	1.515556
6.	4	3.674372	.9615023
7.	4.3	4.10852	.52988
8.	2.7	3.022767	-.9213465
9.	3.6	3.593097	.0188636
10.	4.1	3.589859	1.465039
11.	2.7	2.794364	-.2699653
12.	2.9	2.729938	.4560008
13.	2.5	2.729707	-.6215216
14.	3	3.430356	-1.298388
15.	3.3	4.087624	-2.504239
16.	3.2	3.331268	-.3830797
17.	4.1	3.813245	.7844458
18.	3	2.791182	.5684165
19.	2.6	2.829665	-.6148894
20.	3.7	3.174036	1.515556
21.	4	3.674372	.9615023
22.	4.3	4.10852	.52988
23.	2.7	3.022767	-.9213465
24.	3.6	3.593097	.0188636
25.	4.1	3.589859	1.465039
26.	2.7	2.794364	-.2699653
27.	2.9	2.729938	.4560008
28.	2.5	2.729707	-.6215216
29.	3	3.430356	-1.298388
30.	3.3	4.087624	-2.504239

And you might check to see if the residuals and the predicted values are correlated.

```
. corr r p
```

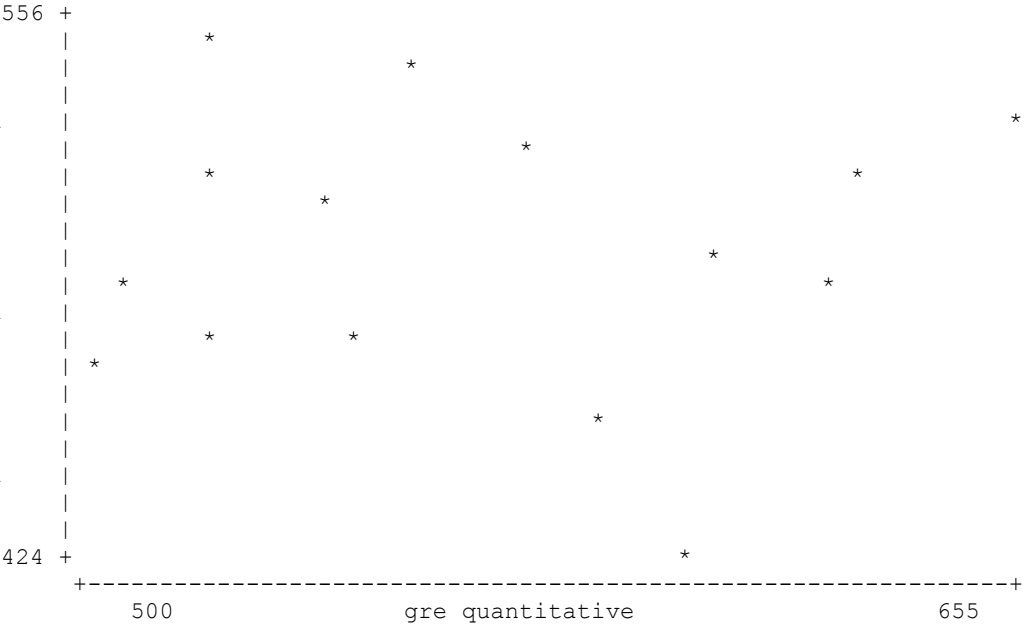
```
(obs=30)

             |        r        p
-------------+------------------
           r |   1.0000
           p |  -0.0313   1.0000
```

It's not a good thing if they are correlated. These are not correlated so things look good.

Then ask what the plot of the residuals look like against the fitted (predicted) values and the independent variables in the model. What follows are the commands and results:

```
plot r p
```

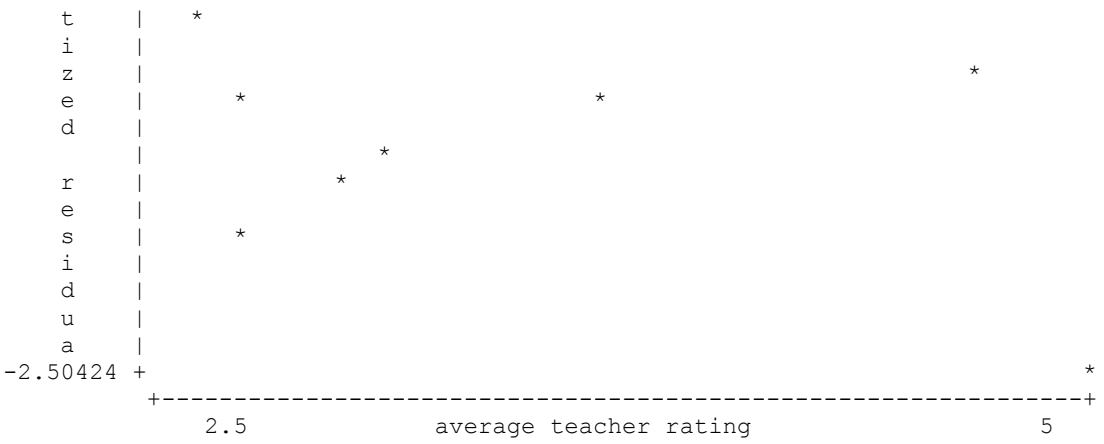

```
plot r greq
```



```
plot r grev

 1.51556 +
    S    |   *
    t    |                                                        *
    u    |
    d    |                *
    e    |                                                    *
    n    | *                                                            *
    t    |           *
    i    |
    z    |                        *
    e    |                 * *
    d    |
         |           *
    r    |      *
    e    |
    s    |                                                           *
    i    |
    d    |
    u    |
    a    |
-2.50424 +                                  *
          +-----------------------------------------------------------------+
              480                    gre verbal                          720
```

```
plot r mat

 1.51556 +
    S    |                                          *
    t    |                                          *
    u    |
    d    |                      *
    e    |                                          *
    n    |                      *                   *
    t    | *
    i    |
    z    |                      *
    e    | *                    *
    d    |
         | *
    r    |                                          *
    e    |
    s    |                      *
    i    |
    d    |
    u    |
    a    |
-2.50424 +                                                             *
          +-----------------------------------------------------------------+
               55             miller's analogies test                    85
```

```
plot r ar

 1.51556 +
    S    |                           *
    t    |                       *
    u    |
    d    |                                        *
    e    |                                             *
    n    | *                                             *
```



IV. Analysis of variance (ANOVA) table for multiple regression (OPTIONAL)

In multiple regression, use the ANOVA table (the upper left hand section of the regression output) to construct another test statistic, the **ANOVA *F* test**, which in this context will allow us to test if all of the regression coefficients (with the exception of the intercept) are simultaneously equal to 0.

We will continue to model the variation of the data as follows:

Total Sum of Squares (SST) = Model Sum of Squares (SSM) + Error Sum of Square (SSE)

Analysis of Variance *F* Test

In the multiple regression model, the hypothesis

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

is tested by the analysis of variance *F* statistic

$$F = \text{MSM/MSE}$$

The *p*-value is the probability that a random variable having the $F(k, n - k - 1)$ distribution is greater than or equal to the calculated value of the *F* statistic.

Table E has the data needed to look up values for the *F* distribution. The degrees of freedom in the numerator (*df* for MSM) is *k*. The degrees of freedom for the denominator (*df* for MSE) is $n - k - 1$.

Note the upper right hand side of the previous regression output. The F is interpreted as the probability of observing a value of greater than the one you got. So the probability of seeing an F-stat greater than 11.13 is virtually zero. The (4, 25) after the letter "F" stands for the degrees of freedom in the numerator and in the denominator. In English, we would reject the null hypothesis that all of the slopes are equal to zero. So the model as a whole appears to be significant. The best thing to do in this case, is to re-run the model with the non significant variables dropped.

```
. regress gpa greq mat
```

```
      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  2,    27) =   18.87
       Model |  6.08344211     2  3.04172105           Prob > F      =  0.0000
    Residual |  4.35122454    27  .161156464           R-squared     =  0.5830
-------------+------------------------------           Adj R-squared =  0.5521
       Total |  10.4346666    29  .359816091           Root MSE      =  .40144

------------------------------------------------------------------------------
         gpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        greq |   .0059763    .001591     3.76   0.001     .0027118    .0092408
         mat |   .0308074   .0083646     3.68   0.001     .0136446    .0479702
       _cons |  -2.129377   .9270377    -2.30   0.030    -4.031501   -.2272527
------------------------------------------------------------------------------
```

What to do now? We note the R-squared has dropped but only marginally (look at the adjusted r-squared). We note that the F statistic got larger (from 11.13 to 18.87) suggesting that this model as a whole is doing a good job. It is a simpler model too.

V. Squared multiple correlation R^2

As was the case for simple linear regression, we have an additional variable that will measure the variation of the response variable that is explained by the multiple regression equation.

R^2 = Model Sum of Squares (SSM) / Total Sum of Squares (SST)

Or

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

Note that for multiple regression, the correlation coefficient r is not very meaningful, since it measures the correlation between only TWO variables, rather than among multiple variables. Thus the "r^2" for multiple regression (now denoted as R^2) is not just the square of r as it was for simple linear regression. It is still, however, a measure that shows the proportion of the total variation in y that is explained by the multiple regression equation (the model).

Some people call R^2 the coefficient of multiple determination. Again, it's just SSM/SST and it is a measure of the percentage of variation in the dependent variable y, that can be accounted for by all the independent variables x in the model.

If you keep adding variables, even if they are not significant, your R-squared will increase. This is a bad practice. Ideally you want to account for as much variation in y as possible with as small a set of variables as possible. So many people look at the ADJUSTED R-SQUARED in multiple regression. This is just the R-squared adjusted to reflect the number of variables in your model. It takes into account the decrease in the degrees of freedom.